The random greedy algorithm for sum-free subsets of $\mathbb{Z}_{2n}$

Patrick Bennett*

Abstract

$S \subseteq \mathbb{Z}_{2n}$ is said to be sum-free if S has no solution to the equation $a + b = c$. The sum-free process on $\mathbb{Z}_{2n}$ starts with $S := \emptyset$, and iteratively inserts elements of $\mathbb{Z}_{2n}$, where each inserted element is chosen uniformly at random from the set of all elements that could be inserted while maintaining that S is sum-free. We prove a lower bound (which holds with high probability) on the final size of S, which matches a more general result of Bennett and Bohman ([2]), and also matches the order of a sharp threshold result proved by Balogh, Morris and Samotij ([1]). We also show that the set S produced by the process has a particular non-pseudorandom property, which is in contrast with several known results about the random greedy independent set process on hypergraphs.

1 Introduction

Let $\mathcal{H}$ be a hypergraph. A set of vertices $S \subset V(H)$ is called *independent* if S contains no edge of $\mathcal{H}$. The random greedy algorithm for independent sets starts with $S = \emptyset$, and then randomly inserts elements into S so long as S remains independent. Specifically, at the i^{th} step, we put $S := S \cup \{v_i\}$ where v_i is chosen uniformly at random from all vertices v such that $S \cup \{v\}$ is independent (we halt when there are no such vertices v). The algorithm terminates with a maximal independent set.

One notable instance of this algorithm is the H-free process for any fixed k-uniform hypergraph H (where $k \geq 2$ so H might be a graph). The process starts with an empty hypergraph on vertex set $[n]$, and iteratively inserts randomly chosen hyperedges so long as we never create a subhypergraph isomorphic to H. The H-free process can be viewed as an instance of the random greedy independent set algorithm running on a hypergraph $\mathcal{H}$ with vertex set $\binom{[n]}{k}$, where each edge of the hypergraph $\mathcal{H}$ corresponds to a subset of $\binom{[n]}{k}$ forming a copy of H.

Bohman [3] analyzed the H-free process when H is the graph K_3, determining (up to a constant) how long the process lasts , and bounding the independence number of the final graph formed. Bohman's analysis of the process included information about the independence number of the graph produced, and gave a second proof that the Ramsey number $R(3,t)$ is $\Omega\left(\frac{t^2}{\log t}\right)$. In the same paper, Bohman considered the K_4-free process and improved the best known lower bound on $R(4,t)$. Bohman and Keevash [5] went on to analyze the H-free process for many other graphs H, resulting in new lower bounds on $R(s,t)$ for fixed $s \geq 5$, and new lower bounds on the Turán numbers of certain bipartite graphs. More recently, Bohman and Keevash [6], and independently Fiz Pontiveros, Griffiths, and Morris [8] proved that the K_3-free process terminates with $\left(\frac{1}{2\sqrt{2}} + o(1)\right) \log^{1/2} n \cdot n^{3/2}$ edges (and also gave

*Computer Science Department, University of Toronto, Toronto, ON, Canada, M5S 3G4. Email: patrickb@cs.toronto.edu. Research supported in part by .

better bounds on the indpendence number of the graph produced by the K_3-free process). We make a conjecture that would generalize this result. With Bohman, the author in [2] consider the independent set process on a general class of hypergraphs, and proved a lower bound (on the size of the final independent set) which generalizes many of the known lower bounds for specific hypergraphs. The general bound in [2] applies to some instances of the H-free process where H is a hypergraph, as well as the k-AP-free process (which chooses elements of $\mathbb{Z}_n$ while avoiding a k-term arithmetic progression).

Theorem 1.1. *Let r and $\epsilon > 0$ be fixed. Let $\mathcal{H}$ be a r-uniform, D-regular hypergraph on N vertices such that $D > N^{\epsilon}$. Define the* degree *of a set $A \subset V(\mathcal{H})$ to be the number of edges of $\mathcal{H}$ that contain A. For $a = 2, \ldots, r-1$ we define $\Delta_a(\mathcal{H})$ to be the maximum degree of A over $A \in \binom{V}{a}$. We also define the* b-codegree *of a pair of distinct vertices v, v' to be the number of edges $e, e' \in \mathcal{H}$ such that $v \in e \setminus e', v' \in e' \setminus e$ and $|e \cap e'| = b$. We let $\Gamma_b(\mathcal{H})$ be the maximum b-codegree of $\mathcal{H}$.*

If

$$\Delta_\ell(\mathcal{H}) < D^{\frac{r-\ell}{r-1}-\epsilon} \qquad \text{for } \ell = 2, \ldots, r-1 \tag{1}$$

and

$$\Gamma_{r-1}(\mathcal{H}) < D^{1-\epsilon} \tag{2}$$

then the random greedy independent set algorithm produces an independent set S in $\mathcal{H}$ with

$$|S| = \Omega\left(N \cdot \left(\frac{\log N}{D}\right)^{\frac{1}{r-1}}\right) \tag{3}$$

with probability $1 - \exp\left\{-N^{\Omega(1)}\right\}$.

The indpendent sets produced by the algorithm tend to have pseudorandom properties. For example, the K_3-free process produces a graph whose independence number is roughly the same as it would be in a random graph with the same edge density ([3], [6], [8]), and for any fixed K_3-free graph G, the number of copies of G in the graph produced by the K_3-free process is roughly the same as it would be in a random graph with the same edge density (see [10]). [2] also has a pseudorandom type result for the independent sets produced by the algorithm, which generalized Wolfovitz's result and which they used to bound the Gowers norm of the set produced by the k-AP-free process. We state this result from [2] now.

Theorem 1.2. *Fix s and a s-uniform hypergraph $\mathcal{G}$ on vertex set $V(\mathcal{H})$ (i..e the same vertex set as the hypergraph $\mathcal{H}$). We let $X_{\mathcal{G}}$ be the number of edges in $\mathcal{G}$ that are contained in the independent set produced at the i^{th} step of the random greedy process on $\mathcal{H}$. Set $p = p(i) = i/N$ and let $i_{\max}$ be the lower bound (3) on the size of the independent set given by the random greedy algorithm given in Theorem 1.1. If no edge of $\mathcal{G}$ contains an edge of $\mathcal{H}$, $i < i_{\max}$ is fixed, $|\mathcal{G}|p^s \to \infty$ and $\Delta_a(\mathcal{G}) = o(p^a|\mathcal{G}|)$ for $a = 1, \ldots, s-1$ then*

$$X_{\mathcal{G}} = |\mathcal{G}|p^s(1 + o(1)).$$

with high probability.

This paper addresses the sum-free process. In this process we look for a set $S \subset \mathbb{Z}_{2n}$ such that S has no solutions to the equation $a + b = c$. Define our edge set E to be the family of all solutions $\{a, b, c\}$ to $a + b = c$ (Such edges $\{a, b, c\}$ are called *Schur triples*). Note that $\{a, b, c\}$ may have $1, 2$, or 3 distinct elements. We write the generic form $\{a, b, c\}$ with the understanding that if, say, $b = c$ then we mean the set $\{a, b\}$ and not a multiset with two copies of b. Thus, S is sum-free if and only if S is an independent set in the hypergraph $\mathcal{H}$

with vertex set $\mathbb{Z}_{2n}$ and edge set E. $\mathcal{H}$ is not uniform but almost all of the edges have size 3. Indeed, each vertex $v \neq 0$ is in $O(1)$ edges of size 2 and no edge of size 1.

Theorem 1.1 cannot be applied directly to the sum-free process, since for example $\mathcal{H}$ is not quite uniform, and also not quite regular. However it is nearly uniform and regular, and these issues alone would not present much difficulty in the analysis. There is a more important way in which $\mathcal{H}$ fails to satisfy the hypotheses of Theorem 1.1: the codegree condition (2). Consider the vertices v and $-v$ for $v \neq 0, n$. Each of $v, -v$ is in about $D = \Theta(n)$ edges, but the 2-codegree of $v, -v$ is also $\Theta(n)$ since whenever we have the equation $v + b = c$ we also have the equation $-v + c = b$. Thus, one of the points of interest in this paper is to see how to overcome this issue.

The largest sum-free subset of $\mathbb{Z}_{2n}$ is the set consisting of the odd numbers. Indeed, Balogh, Morris, and Samotij in [1] proved that $p = \left(\frac{\log n}{3n}\right)^{1/2}$ is a sharp threshold above which we have the following: if G_p is a random subset of $\mathbb{Z}_{2n}$ where each element is included with probability p independently, then w.h.p. the largest sum-free subset of G_p is simply the set of its odd elements. Roughly speaking, the fact that *odd + odd = even* imposes a very specific structure on maximum sum-free subsets of random sets in $\mathbb{Z}_{2n}$. However, this special structure ceases to be relevant for very sparse random sets in $\mathbb{Z}_{2n}$.

Our main theorem is as follows:

Theorem 1.3. *With high probability, the sum-free process run on $\mathbb{Z}_{2n}$ produces a set of size at least*

$$i_0 := \frac{1}{\sqrt{3}} \left(1 - 20 \cdot \frac{\log \log n}{\log n}\right) n^{\frac{1}{2}} \log^{\frac{1}{2}} n$$

We conjecture that a matching (up to a constant factor) upper bound holds, since intuitively the Sum-free process should not produce sets that have density larger than the threshold in [1].

Another point of interest in this paper is to demonstrate a non-pseudorandom statistic of the set S produced by the sum-free process (in contrast to all the pseudorandom results concerning other instances of the hypergraph independent set process). It turns out that the set S tends to contain a large number of pairs $v, -v$, much larger than one would expect in a (uniformly chosen) random set of size $|S|$. Indeed, if $|S| = \Theta(n^{\frac{1}{2}} \log^{\frac{1}{2}} n)$ were chosen uniformly at random, then we would expect S to contain $\Theta(\log n)$ pairs $v, -v$. However, as we will see in the last section of the paper, the number of such pairs in S actually grows like a power of n.

Our analysis of the sum-free process on $\mathbb{Z}_{2n}$ easily extends to $\mathbb{Z}_{2n+1}$. Note that $\mathbb{Z}_{2n+1}$ also has linear-sized sum-free subsets (e.g. the set of all odd numbers less than n). However, since there is no result analogous to [1] for $\mathbb{Z}_{2n+1}$, we work in $\mathbb{Z}_{2n}$.

2 The Algorithm and Associated Random Variables

The random greedy algorithm for sum-free sets starts with a sum-free set $S(0) := \emptyset$, and a set $Q(0)$ of elements that may be inserted into the $S(0)$ without spoiling the sum-free property. Since $0 + 0 = 0$, the element 0 cannot be part of any sum-free set. Any singleton except for $\{0\}$ is sum-free, so $Q(0) = \mathbb{Z}_{2n} \setminus \{0\}$. At step i, the algorithm selects an element $s(i)$ uniformly at random from $Q(i)$ and puts $S(i + 1) := S(i) \cup \{s(i)\}$. The algorithm then determines $Q(i+1)$, the set of elements that could potentially be inserted in $S(i+1)$ without spoiling the sum-free property. If $Q(i+1)$ is empty, the algorithm terminates with a maximal sum-free set.

We call the elements of $S(i)$ *chosen*, the elements of $Q(i)$ *open*, and all other elements of $\mathbb{Z}_{2n}$ *closed*. As we noted above, at the start of the algorithm 0 is closed and every other element is open.

For $k = 2, 3$ we define $E_k(i)$ to be the set of edges $e \subseteq Q(i) \cup S(i)$ such that $|e \cap Q(i)| = k$. We will also define several more random variables which (more or less) represent the degrees of vertices. We will partition the set of edges containing v according to what role v plays in the corresponding equation $a + b = c$. For $v \in \mathbb{Z}_{2n} \setminus S(i) \setminus \{0\}$ and $k = 1, 2, 3$ we define the random variables $D_{k,L}(v, i)$, the set of edges $e \in E$ such that $v \in e$, $e \setminus \{v\} \subseteq Q(i) \cup S(i)$, $|e \cap Q(i) \setminus \{v\}| = k - 1$, and v appears on the left of the equation corresponding to e (i.e. v plays the role of a or b in the equation $a + b = c$); and for $v \in \mathbb{Z}_{2n} \setminus S(i)$, we define $D_{k,R}(v, i)$, the set of edges $e \in E$ such that $v \in e$, $e \setminus \{v\} \subseteq Q(i) \cup S(i)$, $|e \cap Q(i) \setminus \{v\}| = k - 1$ and v is on the right side of the equation. We also define the random variable

$$D_2(v, i) := \{q \in Q(i) : q \neq v, \exists e \in D_{2,L}(v, i) \cup D_{2,R}(v, i). q \in e\}.$$

Many of the variables we just defined take two arguments (a vertex v and the step i). Sometimes for short hand we will suppress the i.

3 Heuristically anticipating the trajectories

We use some heuristics to anticipate the likely values of the random variables throughout the process. Intuitively, we will assume that certain aspects of the hypergraph $\mathcal{H}(i)$ are the same as what they would be if $S(i)$ were a set of i uniformly chosen random elements (rather than having been chosen to satisfy the sum-free property).

Consider a vertex $v \neq 0$, and let's estimate the number of ways v is closed by i random elements. v is in about $3n$ edges of $\mathcal{H}$, and nearly all of those edges have size 3. Thus, we estimate the number of those edges $\{v, x, y\}$ such that both x and y have been chosen as

$$3n \cdot \left(\frac{i}{2n}\right)^2 = \frac{3}{4}\frac{i^2}{n} = \frac{3}{4}t^2$$

where we define

$$t = t(i) := n^{-1/2} i$$

. We heuristically assume that the number of times v is closed is a Poisson random variable, so the probability that v is open should be

$$p = p(t(i)) := e^{-\frac{3}{4}t^2}.$$

Also, we assume that most small sets W of vertices behave independently (i.e. the probability they are all open is $p^{|W|}$). Thus we make the following predictions:

$$Q \approx 2np \qquad E_3 \approx 2n^2p^3 \qquad E_2 \approx 3n^{3/2}tp^2$$

$$D_{3,L}(v) \approx 2np^2 \qquad D_{3,R}(v) \approx np^2 \qquad D_{2,L}(v) \approx 2n^{1/2}tp \qquad D_{2,R}(v) \approx n^{1/2}tp$$

However, there are certain properties of $\mathcal{H}$ (namely the codegrees) that will force us to take a bit more care with certain elements. The elements v and $-v$ are correlated, because $v + b = c \iff -v + c = b$. In other words, $\mathcal{H}$ has a lot of edges $\{v, b, c\}$ that have corresponding edges $\{-v, b, c\}$. Thus, if $-v$ is chosen, then a lot of edges containing v are removed from the hypergraph, in particular $D_{2,L}(v) = 0$. Also, even when $-v \notin S$, there may be many ways to close v that would simultaneously close $-v$, so we do not expect v

and $-v$ to behave independently. With some care, though, we can anticipate what the joint distribution ought to be.

$\mathcal{H}$ has about $2n$ edges $\{v, b, c\}$ with corresponding edges $\{-v, b, c\}$, about n edges $\{v, b, c\}$ without any corresponding edge $\{-v, b, c\}$, and about n edges $\{-v, b, c\}$ without any corresponding edge $\{v, b, c\}$. Thus, the expected number of edges closing either of $v, -v$ is roughly t^2, and the probability that both $v, -v$ are open is $e^{-t^2} = p^{4/3}$. Thus we make a prediction:

$$D_{3,R}(0) \approx np^{4/3}$$

If $-v$ is chosen, then it becomes more likely that v will stay open for longer. Indeed, since $-v$ is chosen we do not have to worry about v ever being closed by any of the corresponding edges containing $-v$. In this situation, v can only be closed by its other n edges, so following a similar calculation the probability that v is open given that $-v$ is chosen ought to be $e^{-\frac{1}{4}t^2} = p^{1/3}$. Thus we predict

$$D_{2,R}(0) \approx n^{1/2} t p^{1/3}$$

Finally we predict the value of $D_{1,R}(0)$. At each step, the variable $D_{1,R}(0)$ either stays the same or increases by 1. The probability of increasing is

$$\frac{D_{2,R}(0)}{Q} \approx \frac{1}{2} n^{-\frac{1}{2}} t p^{-\frac{2}{3}}$$

integrating the above expression gives the prediction:

$$D_{1,R}(0) \approx \frac{1}{2}\left(p^{-\frac{2}{3}} - 1\right).$$

The variables $C(v)$ for $v \neq 0$ will be much smaller, and we will prove that later.

We finish this section with a general conjecture about the random greedy algorithm for independent sets in hypergraphs. This conjecture ought to apply to all sufficiently "nice" (almost uniform, almost regular, not too sparse) hypergraphs. In general, the algorithm should terminate soon after the number of open elements is negligible compared to the number we've already chosen. Our conjecture is that we can predict when this happens using the heuristically derived trajectory for Q.

Conjecture 3.1. *The step i at which the random greedy independent set algorithm terminates w.h.p. is asymptotically the value i such that $Q \approx i$. In particular, the sum-free process terminates when $2np \approx n^{1/2} t$, so when*

$$i = \left(\sqrt{\frac{2}{3}} + o(1)\right) n^{\frac{1}{2}} \log^{\frac{1}{2}} n$$

Even proving the lower bound of the conjecture presents considerable difficulty, particularly due to the fact that we suspect $D_{1,R}(0)$ becomes larger than $D_2(v)$ by that time.

4 Proof Overview

We appeal to the usual differential equations method to establish dynamic concentration of the random variables around the trajectories we heuristically derived. See [12] for an introduction to the standard method. We augment the standard method so as to take advantage of the self-correcting nature of some of the variables. For each variable V and each bound

(i.e. upper and lower) we introduce a *critical interval* $I_V = [a_V, b_V]$. This interval varies with time and has one endpoint at the bound we are trying to establish with the other slightly closer to the expected trajectory. We only track V if and when it enters a critical interval. If V enters the critical interval at step j we 'start' observing a sequence of random variables that is designed to be a sub- or supermartingale with the property that if V eventually passes all the way through the interval (and thereby violates the bound in question) then this martingale has a large variation. When working with the lower bounds we consider the sequence $V - a_V$. This sequence should be a submartingale with initial value (i.e. value at step j) roughly $b_V - a_V$. Note that this sequence becomes negative when the bound in question is violated. Similarly, when working with the upper bound we consider the sequence $b_V - V$, which should be a supermartingale. The event that V ever violates one of the stated bounds is then the union over all 'starting' points j of the event that one of the martingales that start at this point has a large variation. We prove Theorem 1.3 by an application of the union bound, taking the union over all variables V and all 'starting' points for both the upper and lower bounds.

The reason that we focus our attention on these critical intervals is the fact that the expected one-step changes in the variables we consider have self-correcting terms. These terms introduce a drift back toward the expected trajectory when V is far from the expected trajectory. By restricting our attention to the critical intervals we make full use of these terms. In some cases, we manage to prove error bounds that actually *decrease* as the process evolves (i.e. we prove that the variable stays within a window of decreasing width) See [11] and [7] for early applications of this self-correcting phenomenon in applications of the differential equations method for proving dynamic concentration. As we noted above, the critical interval method we use here was introduced in [4].

Another important feature of the proof is that we track some "global" variables, namely E_2 and E_3. One can produce a shorter proof of Theorem 1.3, for some i_0 that is smaller by a constant factor, without tracking E_2 or E_3. We now (very roughly) describe how tracking E_2, E_3 benefits our analysis. We use our control over the local variables (i.e. variables of the form $D_{-,-}(v)$) and Lemma 5.2 to get amplified control over the global variables. This helps us track Q since the expected one-step change in Q can be written in terms of E_2. Doing so is much better than writing the expected one-step change in Q in terms of the local variables, which have larger relative errors than the global ones.

For an arbitrary random variable V we define

$$\Delta V(i) = V(i+1) - V(i).$$

We let $\mathcal{F}_i$ be the filtration of the probability space given by the first i edges chosen by the random greedy matching process.

Define $i_0 := \frac{1}{\sqrt{3}}\left(1 - 20 \cdot \frac{\log\log n}{\log n}\right) n^{\frac{1}{2}} \log^{\frac{1}{2}} n$. Note that we have

$$p(t(i_0)) = n^{-1/4} \log^{10+o(1)} n.$$

Define the stopping time T as the minimum of i_0 and the first step i that any of the following bounds fail:

$$\left|D_{3,L}(v) - 2np^2\right| \le 2f_{d3} \quad \forall v \notin \pm S(i) \cup \{0\} \qquad \left|D_{3,R}(v) - np^2\right| \le f_{d3} \quad \forall v \notin S(i) \cup \{0\}$$

$$\left|D_{2,L}(v) - 2n^{\frac{1}{2}}tp\right| \le 2f_{d2} \quad \forall v \notin \pm S(i) \cup \{0\} \qquad \left|D_{2,R}(v) - n^{\frac{1}{2}}tp\right| \le f_{d2} \quad \forall v \notin \pm S(i) \cup \{0\}$$

$$\left|D_{3,R}(0) - np^{4/3}\right| \le f_{d3,0} \qquad \left|D_{2,R}(0) - n^{1/2}tp^{1/3}\right| \le f_{d2,0}$$

$$\left|E_3 - \tfrac{1}{4}Q^3n^{-1}\right| \le f_{e3} \qquad \left|E_2 - \frac{3}{4}Q^2n^{-\frac{1}{2}}t\right| \le f_{e2}$$

$$D_{1,L}(v), D_{1,R}(v) \le \log^2 n \quad \forall v \ne 0 \qquad \left|D_{1,R}(0) - \frac{1}{2}\left(p^{-\frac{2}{3}} - 1\right)\right| \le f_{d1,0}$$

$$|Q - 2np| \le f_q$$

Theorem 1.3 is proved in sections 6-11. We will show that, for specific choices of the error functions f, that the stopping time $T = i_0$ with high probability. This dynamic concentration result will in turn imply that the algorithm produces a set of size at least i_0 with high probability. As mentioned above, for most of the variables we use critical intervals to prove the error bounds. So these variables V each have two functions f_V and g_V. f_V is the farthest V can ever be from its trajectory, and the width of the critical interval is $f_V - g_V$. In other words, we start paying attention to V only when its distance from its trajectory becomes as large as g_V. These error functions must be carefully chosen to satisfy certain inequalities ("variation equations"), which arise from calculations in subsequent sections. If the reader wishes to look ahead at the variation equations, they are on lines (7), (9), (11), (13), (15), (17), (18), and (19).

$$f_{d2} = n^{\frac{1}{4}}\log^3 n\left(1 + 3t + 2t^2 + 3t^3\right) \qquad g_{d2} = f_{d2} - n^{\frac{1}{4}}\log^3 n$$

$$f_{d2,0} = p^{-\frac{2}{3}}n^{\frac{1}{4}}\log^3 n\left(1 + 2t + 2t^2 + 2t^3\right) \qquad g_{d2,0} = f_{d2,0} - p^{-\frac{2}{3}}n^{\frac{1}{4}}\log^3 n$$

$$f_{d3} = pn^{\frac{3}{4}}\log^3 n\left(1 + 4t + 7t^2\right) \qquad g_{d3} = f_{d3} - pn^{\frac{3}{4}}\log^3 n$$

$$f_{d3,0} = n^{\frac{3}{4}}\log^3 n\left(1 + 3t + 5t^2\right)p^{\frac{1}{3}} \qquad g_{d3,0} = f_{d3,0} - p^{\frac{1}{3}}n^{\frac{3}{4}}\log^3 n$$

$$f_{e2} = n\log^6 n\left(1 + 20t + 70t^2 + 130t^3 + 100t^4 + 120t^5\right) \qquad g_{e2} = f_{e2} - n\log^6 n$$

$$f_{e3} = pn^{\frac{3}{2}}\log^6 n\left(1 + 20t + 70t^2 + 160t^3 + 150t^4\right) \qquad g_{e3} = f_{e3} - pn^{\frac{3}{2}}\log^6 n$$

$$f_q = p^{-1}n^{\frac{1}{2}}\log^6 n\left(1 + 2t + 10t^2 + 35t^3 + 40t^4\right) \qquad g_q = f_q - p^{-1}n^{\frac{1}{2}}\log^6 n$$

$$f_{d1,0} = p^{-\frac{5}{3}}n^{-\frac{1}{4}}\log^3 n(2 + 2t + 2t^2) + p^{-\frac{1}{3}}\log n$$

A note on checking the variation equations: they can be verified in a straightforward (though perhaps a bit tedious) manner as follows. On the left side of the inequality, plug in the values of all of the functions. There will be a common power of n, a power of $\log n$ and a power of p that can be factored out of everything, leaving a factor which is just a polynomial in t. That polynomial will have no positive coefficients and so we just bound it using its constant term (which is valid since $t \ge 0$). For example, to verify (11), note that

$$-6n^{-\frac{1}{2}}tg_{d3} + 6pf_{d2} - 2n^{-\frac{1}{2}}f'_{d3} = pn^{\frac{1}{4}}\log^3 n \cdot (-2 - 7t - 3t^3) \le -2pn^{\frac{1}{4}}\log^3 n$$

5 Preliminaries

In this section we prove several bounds that are necessary for our calculations. Most of these bounds are used to justify big-O terms later.

For $v \neq 0$, we anticipate $D_2(v) \approx D_{2,L}(v) + D_{2,R}(v)$. But to make this estimate valid we need to bound the number of $q \in Q$ such that $q \neq v$ and there are two edges $e_1, e_2 \in D_{2,L}(v) \cup D_{2,R}(v)$ such that $q \in e_1, e_2$. Note that v is only in $O(1)$ many edges of size 2, so let us assume $|e_1| = |e_2| = 3$. In other words, for some $s_1, s_2 \in S$ we have $e_j = \{v, q, s_j\}$. There are cases to consider, according to how the corresponding equations are arranged. Without loss of generality we have one of the following cases:

1. $v + s_1 = q$ and $v + q = s_2$. Then $s_1 + s_2 = 2v$ (so we have $\{s_1, s_2\} \in D_{1,R}(2v)$).
2. $v + s_1 = q$ and $q + s_2 = v$. Then $s_1 + s_2 = 0$ (so we have $\{s_1, s_2\} \in D_{1,R}(0)$).

Thus, before stopping time T we have

$$D_2(v) = D_{2,L}(v) + D_{2,R}(v) - O(1 + D_{1,R}(2v) + D_{1,R}(0)) \tag{4}$$

We now show that if $v \neq 0$ then v does not get closed too many times. Note that $D_{1,L}(v)$ only increases in size on steps when we choose an element of $D_{2,L}(v)$, and on such steps $D_{1,L}(v)$ can increase by at most 2. Before the stopping time T, we have that $\frac{D_{2,L}(v)}{Q} \leq 2n^{-\frac{1}{2}}t$. Thus, $C(v,i)$ is stochastically dominated by $2R$ where $R \sim Bi(i, 2n^{-\frac{1}{2}} \log^{\frac{1}{2}} n)$. An application of the Chernoff bound then tells us that R does not get bigger than $\log^2 n$ w.h.p., and thus the stopping time T does not happen due to the condition on $D_{1,L}(v)$. Bounding $D_{1,R}(v)$ is similar.

We now bound the size of $D_2(v_1) \cap D_2(v_2)$, for $v_1, v_2 \neq 0, v_1 \neq \pm v_2$. Again, each of v_1, v_2 is only in $O(1)$ many edges of size 2, so we assume edges have size 3. Then for each $q \in D_2(v_1) \cap D_2(v_2)$ there is a pair $s_1, s_2 \in S$ such that both $\{v_1, s_1, q\}$ and $\{v_2, s_2, q\}$ are in E. There are cases to consider according to how each of the equations is arranged, but in each case we reach one of the following conclusions: $\{v_1 + v_2, s_1, s_2\} \in E$, $\{v_1 - v_2, s_1, s_2\} \in E$, or $\{v_2 - v_1, s_1, s_2\} \in E$. Thus, by our bounds on the sizes of sets of the form $D_{1,L}$ and $D_{1,R}$ we have, for $v_1, v_2 \neq 0, v_1 \neq \pm v_2$ that

$$|D_2(v_1) \cap D_2(v_2)| = O(\log^2 n) \tag{5}$$

To finish this section, we present a couple of lemmas which we will use several times to estimate things. The following lemma will be used to estimate fractions based on estimates of the numerator and denominator.

Lemma 5.1. *For any real numbers $x, y, \epsilon_x, \epsilon_y$, if we have $x, y \neq 0$ and $\left|\frac{\epsilon_x}{x}\right|, \left|\frac{\epsilon_y}{y}\right| \leq \frac{1}{2}$, then*

$$\frac{x + \epsilon_x}{y + \epsilon_y} - \frac{x}{y} = \frac{y\epsilon_x - x\epsilon_y}{y^2} + O\left(\frac{y\epsilon_x\epsilon_y + x\epsilon_y^2}{y^3}\right)$$

Proof.

$$\begin{aligned}
\frac{x+\epsilon_x}{y+\epsilon_y} - \frac{x}{y} &= \frac{x}{y}\left\{\left(1+\frac{\epsilon_x}{x}\right)\cdot\frac{1}{1+\frac{\epsilon_y}{y}} - 1\right\} \\
&= \frac{x}{y}\left\{\left(1+\frac{\epsilon_x}{x}\right)\cdot\left[1-\frac{\epsilon_y}{y}+O\left(\frac{\epsilon_y^2}{y^2}\right)\right] - 1\right\} \\
&= \frac{x}{y}\left\{\frac{\epsilon_x}{x}-\frac{\epsilon_y}{y}+O\left(\frac{\epsilon_x\epsilon_y}{xy}+\frac{\epsilon_y^2}{y^2}\right)\right\} \\
&= \frac{y\epsilon_x - x\epsilon_y}{y^2} + O\left(\frac{y\epsilon_x\epsilon_y + x\epsilon_y^2}{y^3}\right)
\end{aligned}$$

□

The next lemma is used to estimate global parameters (e.g. the total number of edges) based on estimates of local parameters (e.g. vertex degrees).

Lemma 5.2. *Suppose $(x_i)_{i\in I}$ and $(y_i)_{i\in I}$ are real numbers such that $|x_i - x| \le \delta$ and $|y_i - y| < \epsilon$ for all $i \in I$. Then we have*

$$\left|\sum_{i\in I} x_i y_i - \frac{1}{|I|}\left(\sum_{i\in I} x_i\right)\left(\sum_{i\in I} y_i\right)\right| \le 2|I|\delta\epsilon$$

Proof. The triangle inequality gives

$$\left|\sum_{i\in I}(x_i - x)(y_i - y)\right| \le |I|\delta\epsilon.$$

Rearranging this inequality gives

$$\begin{aligned}
\sum_{i\in I} x_i y_i &= x\sum_{i\in I} y_i + y\sum_{i\in I} x_i - |I|xy \pm |I|\delta\epsilon \\
&= \frac{1}{|I|}\left(\sum_{i\in I} x_i\right)\left(\sum_{i\in I} y_i\right) - |I|\left(\frac{1}{|I|}\sum_{i\in I} x_i - x\right)\left(\frac{1}{|I|}\sum_{i\in I} y_i - y\right) \pm |I|\delta\epsilon.
\end{aligned}$$

□

6 Tracking the D_2 variables

In this section we prove dynamic concentration of the D_2-type variables. We start with $D_{2,L}(v)$ for $v \neq 0$.

For any fixed index $j \le i_0$ we define the stopping time T_j as the minimum of the following indices: T, j and the smallest index $i \ge j$ such that $D_{2,L}(v,i)$ is not in the critical interval

$$\left[2n^{\frac{1}{2}}tp + 2g_{d2}, 2n^{\frac{1}{2}}tp + 2f_{d2}\right]$$

We need to estimate $E[\Delta D_{2,L}(v)|\mathcal{F}_i]$ for $j \le i \le T_j$. The one step change $\Delta D_{2,L}(v)$ has both positive and negative contributions. $D_{2,L}(v)$ gains a pair $\{v, q\}$ when we choose

an element q' such that $\{v,q,q'\} \in D_{3,L}(v)$, except in the case where we happen to have $q \in D_2(q')$ (in which case we may conclude that either $2q'$ or $2q$ is in $D_2(v)$). $D_{2,L}(v)$ loses a pair $\{v,b\}$ when b is chosen or closed.

Thus we can put

$$E[\Delta D_{2,L}(v)|\mathcal{F}_i] = \frac{1}{Q}\left\{2D_{3,L}(v) + O(n^{\frac{1}{2}}tp) - \sum_{\{v,q,s\}\in D_{2,L}(v)} D_2(q)\right\}$$

In the sum above, it is intended that $q \in Q, s \in S$, so we are summing all the ways to lose an edge in $\{v,q,s\} \in D_{2,L}(v)$ due to closing q. The big-O term absorbs the error due to: the possibility that we choose some $q \in \{v,q,q'\} \in D_{3,L}(v)$ such that $q' \in D_2(q)$; the possibility of losing $\{v,q,s\}$ due to choosing q; and the effect of the $O(1)$ many edges of size 2 in $D_{2,L}(v)$.

Almost all of the terms in the sum are $D_2(q) \approx D_{2,L}(q) + D_{2,R}(q) \approx 3n^{\frac{1}{2}}tp$, by (4) and our estimates on the variables we're tracking. However, we may have some terms with q such that $-q \in S$, in which case $D_{2,L}(q) = 0$. However, supposing that $\{v,q,s\} \in D_{2,L}(v)$ and $-q \in S$, we conclude (by considering cases) that either $\{v,s,-q\} \in D_{1,R}(v)$ or $\{-v,s,-q\} \in D_{1,R}(-v)$. Thus there are $O(\log^2 n)$ many such terms in the sum. Now applying (4), our estimate for Q, lemma 5.1, and recalling $t = O(\log^{1/2} n)$ we get

$$\begin{aligned}
&\left|E[\Delta D_{2,L}(v)|\mathcal{F}_i] - 2p + \frac{3}{2}n^{-\frac{1}{2}}tD_{2,L}(v)\right| \le 2n^{-1}p^{-1}f_{d3} + 3n^{-\frac{1}{2}}tf_{d2} \\
&\qquad + O\left(n^{-1}\log n f_q + n^{-1}p^{-1}f_{d2}^2 + n^{-\frac{1}{2}}\log^3 n + n^{-\frac{1}{2}}\log n D_{1,R}(0)\right) \\
&= 2n^{-1}p^{-1}f_{d3} + 3n^{-\frac{1}{2}}tf_{d2} + O\left(n^{-\frac{1}{2}}\log^8 np^{-1}\right)
\end{aligned}$$

and if we use the fact that $D_{2,L}(v)$ is in the critical interval, we get

$$\begin{aligned}
E[\Delta D_{2,L}(v)|\mathcal{F}_i] &\le 2p - 3t^2p + 2n^{-1}p^{-1}f_{d3} + 3n^{-\frac{1}{2}}t(f_{d2} - g_{d2}) \\
&\quad + O\left(n^{-\frac{1}{2}}\log^8 np^{-1}\right) \qquad (6)
\end{aligned}$$

For each vertex $v \neq 0$, we define the sequence of random variables

$$D_{2,L}^{+}(v,i) := \begin{cases} D_{2,L}(v,i) - 2n^{\frac{1}{2}}tp - 2f_{d2} & : \pm v \notin S(i) \\ D_{2,L}^{+}(v,i-1) & : otherwise \end{cases}$$

We will show that the sequence of random variables $D_{2,L}^{+}(v,j)\dots D_{2,L}^{+}(v,T_j)$ is a supermartingale, and then use a deviation inequality to show that w.h.p. $D_{2,L}^{+}(v,i)$ is never positive (and hence $D_{2,L}(v,i)$ does not violate its upper bound). For $j \le i < T_j$

$$\begin{aligned}
E[\Delta D_{2,L}^{+}(v)|\mathcal{F}_i] &\le 2n^{-1}p^{-1}f_{d3} + 3n^{-\frac{1}{2}}t(f_{d2} - g_{d2}) - 2n^{-\frac{1}{2}}f_{d2}' \\
&\quad + O\left(n^{-\frac{1}{2}}\log^8 np^{-1} + n^{-1}f_{d2}''\right) \\
&\le -\Omega\left(n^{-\frac{1}{4}}\log^3 n\right)
\end{aligned}$$

Note that we have used (6), and approximated the 1-step change of deterministic functions using derivatives (i.e. Taylor's theorem). The last line can be verified by observing that

$$2n^{-1}p^{-1}f_{d3} + 3n^{-\frac{1}{2}}t(f_{d2} - g_{d2}) - 2n^{-\frac{1}{2}}f'_{d2} \le -n^{-\frac{1}{4}}\log^3 n \tag{7}$$

Now we use Azuma-Hoeffding to bound the probability that the supermartingale $D^+_{2,L}(v,j)\ldots D^+_{2,L}(v,T_j)$ strays too far.

Lemma 6.1. *Let X_j be a supermartingale, with $|\Delta X_i| \le c_i$ for all i. Then*

$$P(X_m - X_0 \ge a) \le exp\left(-\frac{a^2}{2\sum\limits_{i\le m} c_i^2}\right).$$

We now bound the largest possible 1-step changes in the supermartingale. By examining all the contributions (positive and negative) we see that the largest possible 1-step change is a large negative contribution due to the algorithm choosing some vertex v' such that $D_2(v')$ contains a lot of vertices q in an edge $\{v,q,s\} \in D_{2,L}(v)$. Using (5) to bound the number of such q (for any fixed v') we get

$$\left|D^+_{2,L}(v,i)\right| = O(\log^2 n)$$

Thus, if $D_{2,L}(v,i)$ crosses its upper boundary at the stopping time T, then there is some step j (with $T = T_j$) such that

$$D^+_{2,L}(v,j) \le -2(f_{d2}(t(j)) - g_{d2}(t(j))) + O(\log^2 n)$$

and $D^+_{2,L}(v,T_j) > 0$. Applying Azuma-Hoeffding we see that the event that the supermartingale $D^+_{2,L}(v)$ has such a large upward deviation has probability at most

$$exp\left\{-\Omega\left(\frac{(n^{\frac{1}{4}}\log^3 n)^2}{i_0(\log^2 n)^2}\right)\right\} = o(n^{-2})$$

As there are at most $O(n^{\frac{3}{2}}\log^{\frac{1}{2}} n)$ such supermartingales, with high probability none of them have such a large upward deviation, and $D_{2,L}(v)$ stays below its upper bound for all v.

The lower bound for $D_{2,L}(v)$ is similar, as are the bounds for $D_{2,R}(v)$ for $v \ne 0$.

Now we address $D_{2,R}(0)$. For any fixed index $j \le i_0$ we define the stopping time T_j as the minimum of T, j and the smallest index $i \ge j$ such that $D_{2,R}(0,i)$ is not in the critical interval

$$\left[n^{\frac{1}{2}}tp^{\frac{1}{3}} + g_{d2,0}, n^{\frac{1}{2}}tp^{\frac{1}{3}} + f_{d2,0}\right].$$

Now

$$E[\Delta D_{2,R}(0)|\mathcal{F}_i] = \frac{1}{Q}\left\{2D_{3,R}(0) + O(n^{\frac{1}{2}}tp^{\frac{1}{3}}) - \sum_{\{0,q,s\}\in D_{2,R}(0)} D_2(q)\right\}.$$

and note that for every $\{0,q,s\} \in D_{2,R}(0)$ we have that $-q = s \in S$ and so $D_2(q) = D_{2,R}(q)$. Thus

$$
\begin{aligned}
&\left|E[\Delta D_{2,R}(0)|\mathcal{F}_i] - p^{\frac{1}{3}} + \frac{1}{2}n^{-\frac{1}{2}}tD_{2,R}(0)\right| \\
&\leq n^{-1}p^{-1}f_{d3,0} + \frac{1}{2}n^{-\frac{1}{2}}tp^{-\frac{2}{3}}f_{d2} + O\left(n^{-1}p^{-\frac{2}{3}}\log n f_q + n^{-1}p^{-1}f_{d2,0}f_{d2}\right) \\
&= n^{-1}p^{-1}f_{d3,0} + \frac{1}{2}n^{-\frac{1}{2}}tp^{-\frac{2}{3}}f_{d2} + O\left(n^{-\frac{1}{2}}p^{-\frac{5}{3}}\log^8 n\right) \qquad (8)
\end{aligned}
$$

We define the sequence of random variables

$$D^+_{2,R}(0,i) := D_{2,R}(0,i) - n^{\frac{1}{2}}tp^{\frac{1}{3}} - f_{d2,0}$$

We will show that the sequence $D^+_{2,R}(0,j)\ldots D^+_{2,R}(0,T_j)$ is a supermartingale. For $j \leq i < T_j$

$$
\begin{aligned}
E\left[\Delta D^+_{2,R}(0,i)|\mathcal{F}_i\right] &\leq -\frac{1}{2}n^{-\frac{1}{2}}tg_{d2,0} + n^{-1}p^{-1}f_{d3,0} + \frac{1}{2}n^{-\frac{1}{2}}tp^{-\frac{2}{3}}f_{d2} - n^{-\frac{1}{2}}f'_{d2,0} \\
&\quad + O\left(n^{-\frac{1}{2}}p^{-\frac{5}{3}}\log^8 n + n^{-1}f''_{d2,0}\right) \\
&\leq -\Omega\left(n^{-\frac{1}{4}}\log^3 n \cdot p^{-\frac{2}{3}}\right)
\end{aligned}
$$

Note that in the first line we have used (8) and the fact that $D_{2,R}(0)$ is in the critical interval. The last line can be verified by observing that

$$-\frac{1}{2}n^{-\frac{1}{2}}tg_{d2,0} + n^{-1}p^{-1}f_{d3,0} + \frac{1}{2}n^{-\frac{1}{2}}tp^{-\frac{2}{3}}f_{d2} - n^{-\frac{1}{2}}f'_{d2,0} \leq -p^{-\frac{2}{3}}n^{-\frac{1}{4}}\log^3 n \qquad (9)$$

Again, the biggest possible 1-step changes in the above supermartingale comes from the intersections of D_2 sets. Using (5),

$$\left|\Delta D^+_{2,R}(0,i)\right| = O(\log^2 n)$$

Thus, if $D_{2,R}(0,i)$ crosses its upper boundary at the stopping time T, then there is some step j (with $T = T_j$) such that

$$D^+_{2,R}(0,j) \leq -(f_{d2,0}(t(j)) - g_{d2,0}(t(j))) + \log^2 n$$

and $D^+_{2,R}(0,T_j) > 0$. Applying Azuma-Hoeffding we see that the event that the supermartingale $D^+_{2,R}(0,i)$ has such a large upward deviation has probability

$$exp\left\{-\Omega\left(\frac{(n^{\frac{1}{4}}\log^3 np^{-\frac{2}{3}})^2}{i_0(\log^2 n)^2}\right)\right\} = o(n^{-1}).$$

As there are at most $O(n^{\frac{1}{2}}\log^{\frac{1}{2}} n)$ such supermartingales, with high probability none of them have such large upward deviations, and $D_{2,R}(0)$ stays below its upper bound. Proving the lower bound for $D_{2,R}(0)$ is similar.

7 Tracking the D_3 variables

The variable $D_{3,L}(v)$ is nonincreasing, and we lose a triple $\{v,q,q'\} \in D_{3,L}(v)$ whenever q or q' is closed or chosen. Thus we have

$$E[\Delta D_{3,L}(v)|\mathcal{F}_i] = -\frac{1}{Q}\sum_{\{v,q,q'\}\in D_{3,L}(v)} D_2(q) \cup D_2(q') \cup \{q,q'\}$$

Before the stopping time T, we can estimate the above expression. Most of the terms in the sum will have $-q, -q' \notin S$. Supposing that $\{v, q, q'\} \in D_{3,L}(v)$ and $-q \in S$, we conclude that one of q' or $-q'$ must be in $D_2(v)$. Thus there are $O\left(n^{\frac{1}{2}} tp\right)$ many such terms in the sum. Now applying (4), our control on the variables before the stopping time T, lemma 5.1, and recalling $t = O(\log^{1/2} n)$ we get

$$\begin{aligned}
&\left| E[\Delta D_{3,L}(v) | \mathcal{F}_i] + 3n^{-\frac{1}{2}} t D_{3,L}(v) \right| \leq 6pf_{d2} \\
&\qquad + O\left(pD_{1,R}(0) + p\log^2 n + n^{-\frac{1}{2}} tpf_q + n^{-1}p^{-1} f_{d2} f_{d3}\right) \\
&= 6pf_{d2} + O\left(\log^8 n\right) \qquad (10)
\end{aligned}$$

For a fixed index j, and a fixed element $v \notin \pm S(j)$, we define the sequence of random variables $D^+_{3,L}(v)(j) \ldots D^+_{3,L}(v)(T_j)$, where

$$D^+_{3,L}(v, i) := \begin{cases} D_{3,L}(v, i) - 2np^2 - 2f_{d3} & : v \notin \pm S(i) \\ D^+_{3,L}(v, i-1) & : otherwise \end{cases}$$

and the stopping time T_j is the minimum of T, j and the smallest index i such that $D_{3,L}(v, i)$ is not in the critical interval

$$\left[2np^2 + 2g_{d3}, 2np^2 + 2f_{d3}\right].$$

We will show that the sequences $D_{3,L}(v, j) \ldots D_{3,L}(v, T_j)$ are supermartingales. For $j \leq i < T_j$, we have the inequality

$$\begin{aligned}
&E\left[\Delta D^+_{3,L}(v, i) | \mathcal{F}_i\right] \leq -6n^{-\frac{1}{2}} tg_{d3} + 6pf_{d2} - 2n^{-\frac{1}{2}} f'_{d3} \\
&\qquad + O\left(\log^8 n + n^{-1} f''_{d3}\right) \\
&\leq -\Omega\left(pn^{\frac{1}{4}} \log^3 n\right)
\end{aligned}$$

Note that in the first line we have used (10) and the fact that $D_{3,L}(v)$ is in the critical interval. The last line can be verified by observing that

$$-6n^{-\frac{1}{2}} tg_{d3} + 6pf_{d2} - 2n^{-\frac{1}{2}} f'_{d3} \leq -2pn^{\frac{1}{4}} \log^3 n \qquad (11)$$

Now we use Azuma-Hoeffding to bound the probability that the supermartingale strays too far. By considering the total number of elements that get closed in any one step, we see

$$\left|\Delta D^+_{3,L}(v, i)\right| \leq O\left(n^{\frac{1}{2}} tp\right).$$

Thus, if $D_{3,L}(v, i)$ crosses its upper boundary at the stopping time T, then there is some step j (with $T = T_j$) such that

$$D^+_{3,L}(v, j) \leq -2(f_{d3} - g_{d3}) + O\left(n^{\frac{1}{2}} tp\right)$$

and $D^+_{3,L}(v, T_j) > 0$. Applying Azuma-Hoeffding we see that the event that the supermartingale $D^+_{3,L}(v, i)$ has such a large upward deviation has probability

$$exp\left\{-\Omega\left(\frac{(n^{\frac{3}{4}}\log^3 np)^2}{\sum_{i\le i_0}(n^{\frac{1}{2}}tp)^2}(1+o(1))\right)\right\} = o(n^{-2})$$

As there are at most $O(n^{\frac{3}{2}}\log^{\frac{1}{2}} n)$ such supermartingales, with high probability none of them have such large upward deviations, and $D_{3,L}(v)$ stays below its upper bound for all v.

The lower bound for $D_{3,L}(v)$ is similar, as are the bounds for $D_{3,R}(v)$ for $v \neq 0$.

Now we address $D_{3,R}(0)$. Using the same methods to estimate $E[\Delta D_{3,L}(v)|\mathcal{F}_i]$ before the stopping time T, we get

$$E[\Delta D_{3,R}(0)|\mathcal{F}_i] = -\frac{1}{Q}\sum_{\{0,q,-q\}\in D_{3,R}(0)} D_2(q)\cup D_2(-q)\cup\{q,-q\}$$

but note that each term in the sum has $D_{2,L}(q) = D_{2,L}(-q)$ and so the term is roughly $4n^{1/2}tp$. Thus

$$\begin{aligned}
&\left|E[\Delta D_{3,R}(0)|\mathcal{F}_i] + 2n^{-\frac{1}{2}}tD_{3,R}(0)\right| \\
&\le 2p^{\frac{1}{3}}f_{d2} + O\left(p^{\frac{1}{3}}D_{1,R}(0) + n^{-\frac{1}{2}}p^{\frac{1}{3}}f_q + n^{-1}p^{-1}f_{d2}f_{d3,0}\right) \\
&= 2p^{\frac{1}{3}}f_{d2} + O\left(p^{-\frac{2}{3}}\log^8 n\right) \qquad (12)
\end{aligned}$$

We define the sequence of random variables

$$D^+_{3,R}(0,i) := D_{3,R}(0,i) - n^{\frac{1}{2}}tp^{\frac{4}{3}} - f_{d3,0}.$$

Now for any fixed index $j \le i_0$ we define the stopping time T_j as the minimum of T, j and the smallest index $i \ge j$ such that $D_{3,R}(0,i)$ is not in the critical interval

$$\left[n^{\frac{1}{2}}tp^{\frac{4}{3}} + g_{d3,0}, n^{\frac{1}{2}}tp^{\frac{4}{3}} + f_{d3,0}\right]$$

We will show that the sequence $D^+_{3,R}(0,j)\ldots D^+_{3,R}(0,T_j)$ is a supermartingale. For $j \le i < T_j$

$$\begin{aligned}
E\left[\Delta D^+_{3,R}(0,i)|\mathcal{F}_i\right] &\le -2n^{-\frac{1}{2}}tg_{d3,0} + 2p^{\frac{1}{3}}f_{d2} - n^{-\frac{1}{2}}f'_{d3,0} \\
&\quad + O\left(p^{-\frac{2}{3}}\log^8 n + n^{-1}f''_{d3,0}\right) \\
&\le -\Omega\left(p^{\frac{1}{3}}n^{\frac{1}{4}}\log^3 n\right)
\end{aligned}$$

Note that in the first line we have used (12) and the fact that $D_{3,R}(0)$ is in the critical interval. The last line can be verified by observing that

$$-2n^{-\frac{1}{2}}tg_{d3,0} + 2p^{\frac{1}{3}}f_{d2} - n^{-\frac{1}{2}}f'_{d3,0} \le -p^{\frac{1}{3}}n^{\frac{1}{4}}\log^3 n. \qquad (13)$$

Now we use Azuma-Hoeffding to bound the probability that the supermartingale strays too far. We have

$$\left|\Delta D^+_{3,R}(0,j)\right| \le O\left(n^{\frac{1}{2}}tp\right)$$

Thus, if $D_{3,R}(0,i)$ crosses its upper boundary at the stopping time T, then there is some step j (with $T = T_j$) such that

$$D^+_{3,R}(0,j) \le -(f_{d3,0} - g_{d3,0}) + O\left(n^{\frac{1}{2}}tp\right)$$

and $D^+_{3,R}(0,T_j) > 0$. Applying Azuma-Hoeffding we see that the event that the supermartingale $D^+_{3,R}(0)$ has such a large upward deviation has probability $o(n^{-1})$. As there are at most $O(n^{\frac{1}{2}}\log^{\frac{1}{2}} n)$ such supermartingales, with high probability none of them have such large upward deviations, and $D_{3,R}(0)$ stays below its upper bound. The lower bound for $D_{3,R}(0)$ is similar.

8 Tracking the E_2 variable

We have the formula

$$\begin{aligned}E[\Delta E_2|\mathcal{F}_i] &= \frac{1}{Q}\sum_{q\in Q}\left[D_{3,L}(q) + D_{3,R}(q) + O(n^{\frac{1}{2}}tp) - \sum_{q'\in D_2(q)} D_{2,L}(q') + D_{2,R}(q')\right]\\ &= \frac{1}{Q}\left[3E_3 - \sum_{q\in Q} D_2(q)^2\right] + O(pn^{\frac{1}{2}}\log n + \log^2 n + p^{-\frac{2}{3}})\end{aligned}$$

We apply (5.2) to the sum $\sum_{q\in Q} D_2(q)^2$, except we have to be careful because there are a few terms that are significantly smaller than others, namely the terms corresponding to $q \in -S(i)$. Using simple bounds on the number of such terms, we have

$$\sum_{q\in Q} D_2(q)^2 \in \frac{4E_2^2}{Q} \pm 18Qf_{d2}^2 + O\left(n^{\frac{3}{2}}t^3p^2\right)$$

and

$$\left|E[\Delta E_2|\mathcal{F}_i] - \frac{3E_3}{Q} + \frac{4E_2^2}{Q^2}\right| \le 18f_{d2}^2 + O\left(pn^{\frac{1}{2}}\log^2 n + \log^2 n + p^{-\frac{2}{3}}\right) \tag{14}$$

For a fixed index j, we define the sequence of random variables $E_2^+(j)\dots E_2^+(T_j)$, where

$$E_2^+(i) := E_2(i) - \frac{3}{4}Q(i)^2 n^{-\frac{1}{2}}t - f_{e2}$$

and the stopping time T_j is the minimum of T, j and the smallest index i such that E_2 is not in the critical interval

$$\left[\frac{3}{4}Q(i)^2n^{-\frac{1}{2}}t + g_{e2}, \frac{3}{4}Q(i)^2n^{-\frac{1}{2}}t + f_{e2}\right]$$

We will choose the functions g_{e2}, f_{e2} so that these are supermartingales. We have the inequality

$$
\begin{aligned}
&E\left[\Delta\left(E_2^+(i)\right)|\mathcal{F}_i\right] \\
&\le \frac{3E_3}{Q} - \frac{4E_2^2}{Q^2} + 18f_{d2}^2 - \frac{3}{4}n^{-1}Q^2 - \frac{3}{4}n^{-\frac{1}{2}}tQ\left(-\frac{2E_2}{Q}\right) - n^{-\frac{1}{2}}f'_{e2} \\
&\qquad +O\left(pn^{\frac{1}{2}}\log^2 n + \log^2 n + p^{-\frac{2}{3}} + n^{-1}f''_{e2}\right) \\
&= \frac{3}{Q}\left(E_3 - \frac{1}{4}n^{-1}Q^3\right) - \frac{4E_2}{Q^2}\left(E_2 - \frac{3}{4}n^{-\frac{1}{2}}tQ^2\right) + 18f_{d2}^2 - n^{-\frac{1}{2}}f'_{e2} \\
&\qquad +O\left(pn^{\frac{1}{2}}\log^2 n + \log^2 n + p^{-\frac{2}{3}} + n^{-1}f''_{e2}\right) \\
&\le \frac{3}{2}n^{-1}p^{-1}f_{e3} - 3tn^{-\frac{1}{2}}g_{e2} + 18f_{d2}^2 - n^{-\frac{1}{2}}f'_{e2} \\
&\qquad +O\left(pn^{\frac{1}{2}}\log^2 n + \log^2 n + p^{-\frac{2}{3}} + n^{-1}f''_{e2} + n^{-2}p^{-2}f_{e3}f_q + n^{-2}p^{-2}f_{e2}^2 + n^{-\frac{3}{2}}tp^{-2}f_qf_{e2}\right) \\
&\le -\Omega\left(n^{\frac{1}{2}}\log^6 n\right)
\end{aligned}
$$

Note that in the first line we have used (14). The second line sheds some light on the reason why we track E_2 in terms of Q instead of a completely deterministic function. By doing so, we arrive at a supermartingale calculation in which main terms neatly group together and cancel, leaving only a few small error terms behind. The last line can be verified by observing that

$$
\frac{3}{2}n^{-1}p^{-1}f_{e3} - 3tn^{-\frac{1}{2}}g_{e2} + 18f_{d2}^2 - n^{-\frac{1}{2}}f'_{e2} \le -\frac{1}{2}n^{\frac{1}{2}}\log^6 n \tag{15}
$$

Now we use Azuma-Hoeffding to bound the probability that the supermartingale strays too far. We have

$$
\left|\Delta E_2^+(i_0, j)\right| \le O(n^{\frac{1}{2}}tpf_{d2}) = O(n^{\frac{3}{4}}\log^5 n \cdot p)
$$

Thus, if $E_2(i)$ crosses its upper boundary at the stopping time T, then there is some step j (with $T = T_j$) such that

$$
E_2^+(j) \le -(f_{e2} - g_{e2}) + O(n^{\frac{3}{4}}\log^5 n \cdot p)
$$

and $E_2^+(T_j) > 0$. Applying Azuma-Hoeffding we see that the event that the supermartingale E_2^+ has such a large upward deviation has probability at most

$$
exp\left\{-\Omega\left(\frac{(n\log^6 n)^2}{\sum_{i\le i_0}(n^{\frac{3}{4}}\log^5 n \cdot p)^2}\right)\right\} = o(n^{-1})
$$

As there are at most $O(n^{\frac{1}{2}}\log^{\frac{1}{2}} n)$ such supermartingales, with high probability none of them have such large upward deviations, and E_2 stays below its upper bound. The lower bound for E_2 is similar.

9 Tracking the E_3 variable

We have the formula

$$
\begin{aligned}
E[\Delta E_3|\mathcal{F}_i] &= -\frac{1}{Q}\sum_{q\in Q}\left[D_{3,L}(q)+D_{3,R}(q)+\sum_{q'\in D_2(q)}(D_{3,L}(q')+D_{3,R}(q'))\right]\\
&= -\frac{1}{Q}\sum_{q\in Q} D_2(q)\cdot(D_{3,L}(q)+D_{3,R}(q))+O(np^2)
\end{aligned}
$$

We will apply (5.2) to the sum $\sum_{v\in Q} D_2(v)\cdot D_3(v)$, and again we must be careful because a few of the terms are smaller than others. We arrive at

$$\sum_{v\in Q} D_2(v)\cdot D_3(v) \in \frac{6E_3E_2}{Q} \pm 18Qf_{d2}f_{d3} + O\left(n^2t^2p^3\right)$$

And so

$$\left|E[\Delta E_3|\mathcal{F}_i] + \frac{6E_3E_2}{Q^2}\right| \le 18f_{d2}f_{d3} + O\left(np^2\log n\right) \tag{16}$$

For a fixed index j, we define the sequence of random variables $E_3^+(j)\dots E_3^+(T_j)$, where

$$E_3^+(i) := E_3(i) - \frac{1}{4}Q(i)^3n^{-1} - f_{e3}$$

and the stopping time T_j is the minimum of T, j and the smallest index i such that E_3 is not in the critical interval

$$\left[\frac{1}{4}Q(i)^3n^{-1}+g_{e3}, \frac{1}{4}Q(i)^3n^{-1}+f_{e3}\right].$$

We will choose the function g_{e3}, f_{e3} so that these are supermartingales. We have the inequality

$$
\begin{aligned}
&E\left[\Delta E_3^+(i)|\mathcal{F}_i\right]\\
&\le -\frac{6E_3E_2}{Q^2} + 18f_{d2}f_{d3} - \frac{1}{4}n^{-1}\cdot 3Q^2\left(-\frac{2E_2}{Q}\right) - n^{-\frac{1}{2}}f'_{e3}\\
&\qquad + O\left(np^2\log n + n^{-1}f''_{e3}\right)\\
&\le -\frac{9}{2}n^{-\frac{1}{2}}tg_{e3} + 18f_{d2}f_{d3} - n^{-\frac{1}{2}}f'_{e3}\\
&\qquad + O\left(n^{-2}p^{-2}f_{e3}f_{e2} + n^{-\frac{3}{2}}p^{-1}f_{e3}f_q + np^2\log n + n^{-1}f''_{e3}\right)\\
&\le -\Omega\left(n\log^6 n\cdot p\right)
\end{aligned}
$$

Note that in the first line we have used (16). Similarly to the analogous calculation for E_2, we were able to group main terms together and cancel them. The last line can be verified by observing that

$$-\frac{9}{2}n^{-\frac{1}{2}}tg_{e3} + 18f_{d2}f_{d3} - n^{-\frac{1}{2}}f'_{e3} \le -2pn\log^6 n \tag{17}$$

Now we use Azuma-Hoeffding to bound the probability that the supermartingale strays too far. We have

$$\left|\Delta E_3^+(i)\right| \leq O(np^2 f_{d2} + n^{\frac{1}{2}} tp f_{d3}) = O(n^{\frac{5}{4}} \log^{\frac{9}{2}} n \cdot p^2).$$

Thus, if $E_3(i)$ crosses its upper boundary at the stopping time T, then there is some step j (with $T = T_j$) such that

$$E_3^+(j) \leq -(f_{e3} - g_{e3}) + O(n^{\frac{5}{4}} \log^{\frac{9}{2}} np^2)$$

and $E_3^+(T_j) > 0$. Applying Azuma-Hoeffding we see that the event that the supermartingale E_3^+ has such a large upward deviation has probability at most

$$exp\left\{-\Omega\left(\frac{\left(n^{\frac{3}{2}} \log^6 n \cdot p\right)^2}{\sum_{i \leq i_0} \left(n^{\frac{5}{4}} \log^{\frac{9}{2}} n \cdot p^2\right)^2}\right)\right\} = o(n^{-1})$$

As there are at most $O(n^{\frac{1}{2}} \log^{\frac{1}{2}} n)$ such supermartingales, with high probability none of them have such large upward deviations, and E_3 stays below its upper bound. The lower bound for E_3 is similar.

10 Tracking the Q variable

We have

$$E[\Delta Q | \mathcal{F}_i] = -1 - \frac{1}{Q} \sum_{q \in Q} D_2(q) = -\frac{2E_2}{Q} + O(\log^2 n + p^{-\frac{2}{3}})$$

For a fixed index j , we define the sequence of random variables $Q^+(j) \ldots Q^+(T_j)$, where

$$Q^+(i) := Q(i) - 2np - f_q$$

and the stopping time T_j is the minimum of T, j and the smallest index i such that Q is not in the critical interval

$$[2np + g_q, 2np + f_q]\,.$$

We will choose the function g_q, f_q so that these are supermartingales. We have the inequality

$$\begin{aligned}
&E\left[\Delta\left(Q - 2np - f_q\right) | \mathcal{F}_i\right] \leq -\frac{2E_2}{Q} + 3n^{\frac{1}{2}} tp - n^{-\frac{1}{2}} f_q' + O(\log^2 n + p^{-\frac{2}{3}} + n^{-1} f_q'') \\
&\leq -\frac{3}{2} Q n^{-\frac{1}{2}} t + \frac{2f_{e2}}{Q} + 3n^{\frac{1}{2}} tp - n^{-\frac{1}{2}} f_q' + O(\log^2 n + p^{-\frac{2}{3}} + n^{-1} f_q'') \\
&\leq n^{-1} p^{-1} f_{e2} - \frac{3}{2} n^{-\frac{1}{2}} t g_q - n^{-\frac{1}{2}} f_q' + O\left(\frac{f_{e2} f_q}{n^2 p^2} + \log^2 n + p^{-\frac{2}{3}} + n^{-1} f_q''\right) \\
&\leq -\Omega\left(\log^6 n \cdot p^{-1}\right)
\end{aligned}$$

Note that in the first line we have used (10). The last line can be verified by observing that

$$n^{-1} p^{-1} f_{e2} - \frac{3}{2} n^{-\frac{1}{2}} t g_q - n^{-\frac{1}{2}} f_q' \leq -p^{-1} \log^6 n. \tag{18}$$

Now we use Azuma-Hoeffding to bound the probability that the supermartingale strays too far. We have

$$\left|\Delta Q^{+}(i_0, j)\right| \le O(f_{d2}) = O(n^{\frac{1}{4}} \log^{\frac{9}{2}} n).$$

Thus, if $Q(i)$ crosses its upper boundary at the stopping time T, then there is some step j (with $T = T_j$) such that

$$Q^{+}(j) \le -(f_q - g_q) + O(n^{\frac{1}{4}} \log^{\frac{9}{2}} n)$$

and $Q^{+}(T_j) > 0$. Applying Azuma-Hoeffding we see that the event that the supermartingale Q^{+} has such a large upward deviation has probability

$$exp\left\{-\Omega\left(\frac{(n^{\frac{1}{2}} \log^6 n \cdot p^{-1})^2}{i_0(n^{\frac{1}{4}} \log^{\frac{9}{2}} n)^2}\right)\right\} = o(n^{-1})$$

As there are at most $O(n^{\frac{1}{2}} \log^{\frac{1}{2}} n)$ such supermartingales, with high probability none of them have such large upward deviations, and Q stays below its upper bound. The lower bound for Q is similar.

11 $D_{1,R}(0)$

We have the expected 1-step change

$$E[\Delta D_{1,R}(0)|\mathcal{F}_i] = \frac{D_{2,R}(0)}{Q}$$

and so before T,

$$\begin{aligned}&\left|E[D_{1,R}(0)|\mathcal{F}_i] - n^{-\frac{1}{2}} t p^{-\frac{2}{3}}\right| \\ &\le \frac{1}{2} n^{-1} p^{-1} f_{d2,0} + O\left(n^{-\frac{3}{2}} p^{-\frac{5}{3}} \log n f_q\right)\end{aligned}$$

We define the sequence of random variables

$$C^{+}(0,i) := C(0,i) - \frac{1}{2}\left(p^{-\frac{2}{3}} - 1\right) - h_{d1,0}$$

where

$$h_{d1,0} := n^{-\frac{1}{4}} \log^3 n (2 + 2t + 2t^2) p^{-\frac{5}{3}}.$$

we have

$$\begin{aligned}&E\left[\Delta C^{+}(0,i)|\mathcal{F}_i\right] \\ &\le \frac{1}{2} n^{-1} p^{-1} f_{d2,0} - n^{-\frac{1}{2}} h'_{d1,0} \\ &\qquad + O\left(n^{-\frac{3}{2}} p^{-\frac{5}{3}} \log n f_q + n^{-1} h''_{d1,0}\right) \\ &\le -\Omega\left(n^{-\frac{3}{4}} \log^3 n p^{-\frac{5}{3}}\right)\end{aligned}$$

The last line can be verified by observing that

$$\frac{1}{2}n^{-1}p^{-1}f_{d2,0}-n^{-\frac{1}{2}}h'_{d1,0}\leq-\frac{3}{2}n^{-\frac{3}{4}}\log^3 np^{-\frac{5}{3}}. \tag{19}$$

We will apply the following inequality due to Freedman [9].

Lemma 11.1. *Let X_i be a supermartingale, with $\Delta X_i \leq C$ for all i, and $V(i) := \sum_{k\leq i} Var[\Delta X(k)|\mathcal{F}_k]$ Then*

$$P\left[\exists i : V(i)\leq v, X_i - X_0 \geq d\right] \leq \exp\left(-\frac{d^2}{2(v+Cd)}\right).$$

For our application of this inequality, we can use $C=1$. To determine a suitable value for v, note first that before T we have

$$Var[\Delta C^+(0,k)|\mathcal{F}_k] = Var[\Delta C(0,k)|\mathcal{F}_k] \leq E\left[\left(\Delta C(0,k)\right)^2|\mathcal{F}_k\right]$$
$$=\frac{D_{2,R}(0,k)}{Q(k)} < n^{-\frac{1}{2}}t(k)p(t(k))^{-\frac{2}{3}}$$

so we bound the sum

$$\sum_{k\leq i} n^{-\frac{1}{2}}t(k)p(t(k))^{-\frac{2}{3}} \leq \int_0^{t(i)} \tau p(\tau)^{-\frac{2}{3}}d\tau < p(t(i))^{-\frac{2}{3}}$$

(note that $tp^{-2/3}$ is increasing in t so we may bound the sum with an integral) so we set $v=p(t(i))^{-\frac{2}{3}}$. Now we see by Lemma 11.1 that with $d=p(t(i))^{-\frac{1}{3}}\log n$, with high probability the supermartingale $D^+_{1,R}(0,i)$ is no larger than d. Therefore before T we have the upper bound

$$D_{1,R}(0,i)\leq\frac{1}{2}\left(p^{-\frac{2}{3}}-1\right)+h_{d1,0}+p(t(i))^{-\frac{1}{3}}\log n.$$

The lower bound for $D_{1,R}(0,i)$ is similar.

In particular, $D_{1,R}(0,i_0)=\frac{1}{2}p(t(i_0))^{-\frac{2}{3}}(1+o(1))=\tilde{\Theta}\left(n^{1/6}\right)$, while $D_{1,R}(v,i_0), D_{1,L}(v,i_0) = O\left(\log^2 n\right)$ for all $v\neq 0$. The behavior of $D_{1,R}(0)$, and the fact that $D_{1,R}(0)$ appears in many of our big-O terms, would seem to indicate that some new ideas would be needed to track the sum-free process much further (e.g. to prove Conjecture 3.1).

Acknowledgements: The author would like to thank Tom Bohman for several helpful discussions

References

[1] J. Balogh, R. Morris, W. Samotij. *Random Sum-free Subsets of Abelian Groups*, Israel Journal of Mathematics, 199 (2014) 651-685.

[2] P. Bennett, T. Bohman. *A note on the random greedy independent set algorithm, arxiv: 1308.3732*

[3] T. Bohman, *The Triangle-Free Process*, Advances in Mathematics, 221 (2009) 1653-1677.

[4] T. Bohman, A.M. Frieze and E. Lubetzky, *A note on the random greedy triangle-packing algorithm*, Journal of Combinatorics 1 (2010) 477-488.

[5] T. Bohman, P. Keevash, *The early evolution of the H-free process*. Inventiones Mathematicae, 181 (2010) 291-336

[6] T. Bohman, P. Keevash, *Dynamic concentration of the triangle-free process*, The Seventh European Conference on Combinatorics, Graph Theory and Applications 16 (2013) 489-495.

[7] T. Bohman, M. Picollelli, Evolution of SIR epidemics on random graphs with a fixed degree sequence. *Random Structures and Algorithms* **41** (2012) 179–214.

[8] G. Fiz Pontiveros, S. Griffiths, R. Morris *The triangle-free process and R(3,k), arxiv: 1302.6279*

[9] David A. Freedman, *On Tail Probabilities for Martingales*, The Annals of Probability 3 (1975), 100-118.

[10] G. Wolfovitz, Triangle-free subgraphs in the triangle-free process, *Random Structures Algorithms* **39** (2011), 539–543.

[11] A. Telcs, N. Wormald, S. Zhou, Hamiltonicity of random graphs produced by 2-processes. *Random Structures and Algorithms* **31** (2007) 450–481.

[12] N. Wormald, The differential equation method for random graph processes and greedy algorithms, in *Lectures on Approximation and Randomized Algorithms* (M. Karonski and H.J. Prömel, eds), pp. 73–155. PWN, Warsaw, 1999.